Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Empowering Surgical Embodied Intelligence: A Blockchain-based Federated Learning Framework for Collaborative Video Analysis

Pengpeng Li
Nanjing University of Science and Technology
Xiangbo Shu
shuxb@njust.edu.cn

Nanjing University of Science and Technology
Chun-Mei Feng
University College Dublin
Yifei Feng
Nanjing Medical University
Xin Gao
King Abdullah University of Science and Technology
Jinhui Tang
Nanjing Forestry University

Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11331331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Empowering Surgical Embodied Intelligence: A Blockchain-based Federated Learning Framework for Collaborative Video Analysis

Pengpeng Li[1], Xiangbo Shu[1*], Chun-Mei Feng[2], Yifei Feng[3,4], Xin Gao[5], Jinhui Tang[6*]

1.School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

2.School of Computer Science, University College Dublin (UCD), Dublin, Ireland

3.The First School of Clinical Medicine, Nanjing Medical University, Nanjing, China

4.Department of General Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

5.Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia

6.College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, China

**Corresponding authors: shuxb@njust.edu.cn, tangjh@njfu.edu.cn*

Abstract

Automated surgical video analysis serves as the critical foundation for surgical embodied intelligence. However, privacy regulations and data-sharing restrictions pose significant challenges to its collaborative development in computer-assisted surgery. While federated learning (FL) offers a decentralized approach, it remains vulnerable to privacy breaches and single-point failures due to its reliance on a central server. Here, we propose a blockchain-based federated learning (BFL) framework for secure and collaborative surgical video analysis. BFL integrates blockchain's decentralized consensus and encryption mechanisms with collaborative learning, eliminating the need for a central coordinator. The framework employs cross-validation guided parameter encryption and dynamic weight aggregation to safeguard data privacy, and adopts a hybrid model architecture "shared + private" to balance global generalization with local personalization. We apply BFL to distributed surgical video analysis environments involving both medical center and perspective heterogeneity, using data collected from multiple medical centers. Experimental results show that BFL outperforms traditional federated learning methods in terms of both accuracy and robustness. Overall, by establishing a trustworthy model for collaborative medical AI, our framework facilitates secure data sharing across medical centers, empowering surgical embodied intelligence with the visual perception essential for future autonomous surgical assistance.

1. Introduction

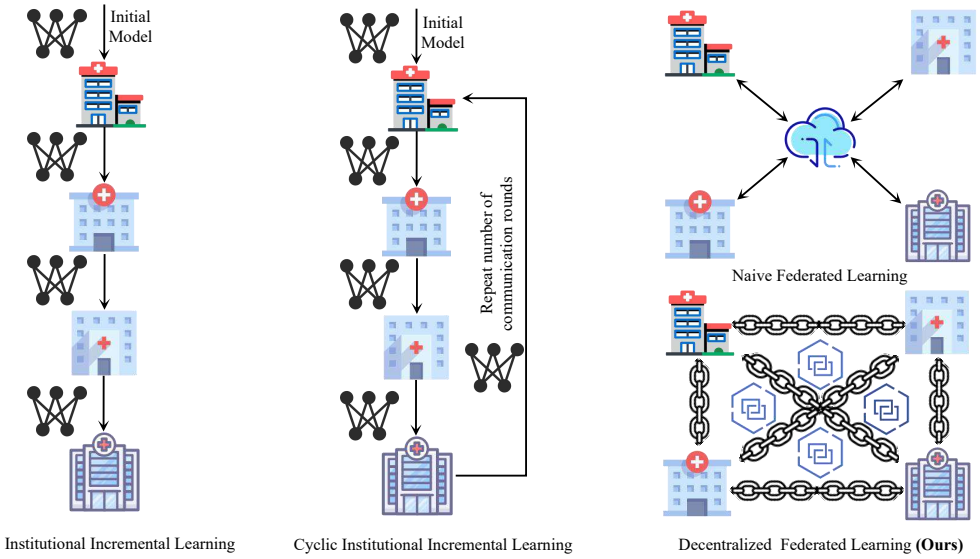
Surgical video analysis is pivotal in medical research, providing not only accurate information on surgical scenes to enhance understanding and decision support, but also serves as the critical perceptual foundation for the emerging field of surgical embodied intelligence [1, 2, 3, 4, 5]. While deep learning has driven significant advancements in tasks like surgical phase recognition [6, 7, 8] and tool detection [9, 10], its broader application is severely constrained by the inherently sensitive and strictly regulated nature of medical data. Surgical videos are typically distributed across independent medical centers, facing stringent privacy regulations that strictly prohibit direct sharing of raw data, thereby leading to isolated data silos. This data fragmentation severely impedes the integration of large-scale, diverse datasets crucial for training robust deep learning models, while manual annotation of surgical videos remains labor-intensive, time-consuming, and highly dependent on domain expertise [11, 12, 13].

In recent years, as shown in Fig. **1**a, Institutional Incremental Learning (IIL) and Cyclic Institutional Incremental Learning (CIIL) have been proposed to address the challenge of distributed data across multiple independent medical centers [14, 15]. IIL adopts a sequential training approach across institutions but faces challenges such as catastrophic forgetting and limited efficiency in utilizing data from all centers simultaneously. CIIL builds on this by enabling multiple rounds of communication between institutions, but the increased communication overhead and latency restrict its scalability for larger datasets. To address these challenges, Federated Learning (FL) has emerged as a promising alternative, offering an efficient and privacy-preserving framework [16, 17, 18]. As also illustrated in Fig. **1**a, Naive Federated Learning (NFL) operates with a central server that first distributes global model parameters to selected clients. Clients then train and update their models based on local data before returning the updated parameters. Subsequently, the server aggregates these parameters using algorithms like FedAvg [19] to refine the global model, which it then redistributes to all participants.

However, the centralized NFL architecture, due to its heavy reliance on a single trusted central server, suffers from a single point of failure if the server is compromised or behaves maliciously. This centralized paradigm conflicts with medical institutions' demands for data autonomy and strict regulatory compliance, as submitting sensitive model parameters to external, unaudited central entities inherently undermines their control and compliance. Unaudited transmission of gradients or parameters, even without raw data, fundamentally fails to meet the rigorous security and auditability required by medical privacy regulations. Compounding this, the inherent unique characteristics of surgical videos, such as diverse perspectives and institutional practices (Fig. **1**b,c), further exacerbate the challenge of balancing global model generalization with local personalization needs [11, 12, 13].

Here we propose BFL, a blockchain-based federated learning framework specifically designed for collaborative surgical video analysis, which effectively addresses the critical unmet needs for security and reliability in multi-institutional medical AI collaboration. By leveraging the immutable, transparent, and decentralized properties of a permissioned blockchain, BFL establishes a trustworthy and auditable environment for model training,

a



b

Dataset	**Cholec80**			**HeiChole**		**m2cai16-workflow**		
Desciption	The Cholec80, HeiChole, and m2cai16 datasets include laparoscopic videos with annotations for phase recognition, etc., aiding the development of computer vision models for real-time and robotic surgery.							
No. of Videos	80			24		41		
Medical Center K	Center 0	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6	Center 7
Heterogeneous types	Perspective Heterogeneity			Institutional Heterogeneity		Perspective Heterogeneity		
Classes	Preparation, Calot Triangle Dissection, Clipping Cutting, Gallbladder Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction							

c

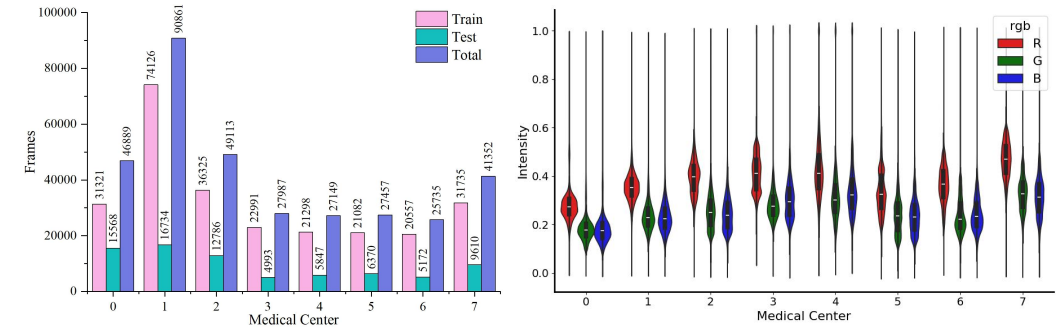


Fig. 1: **Comprehensive schematic**. **a** Learning strategies. Institutional Incremental Learning (sequential), Cyclic Incremental Learning (iterative refinement), and Decentralized Federated Learning (ours, collaborative with privacy preservation). **b** Dataset summary. Cholec80, HeiChole, and m2cai16-workflow datasets with heterogeneous laparoscopic videos across centers (0–7). **c** Data distribution. (left) Frame counts for training, testing, and total per center, (right) RGB intensity distributions showing visual heterogeneity.

thereby fundamentally surpassing traditional centralized federated learning paradigms. The framework incorporates a meticulously designed blockchain-based consensus aggregation protocol, which not only decentralizes aggregation authority but, more importantly, introduces a cross-validation-guided verifiable contribution mechanism to ensure fairness and robustness of global model updates against malicious participants. In addition, BFL employs privacy-preserving homomorphic encryption for secure parameter exchange, enabling computations directly on encrypted shared model parameters and ensuring that sensitive updates comply with stringent medical privacy standards. Furthermore, in response to the dual need for universal surgical knowledge and institution-specific adaptation in heterogeneous medical data (Fig. 1b,c), BFL features a hybrid "shared + private" model architecture that achieves a unique and optimized balance between global generalization and local personalization.

BFL demonstrates enhanced robustness and generalizability, achieving superior performance across key metrics (accuracy, F1 score, Jaccard index, and recall) compared to mainstream FL methods and maintaining stable performance on diverse surgical video datasets. Comprehensive evaluations validate our framework's exceptional robustness against institutional heterogeneity, visual variability, and malicious participants. Beyond these technical advancements, BFL offers significant clinical value by enabling genuine collaborative model training across medical centers without requiring costly raw data exchange. Crucially, by establishing a trustworthy and scalable distributed foundation model, the framework facilitates the secure aggregation of diverse surgical scenarios. This aggregation serves as a prerequisite for developing surgical embodied intelligence agents, thereby providing support for the transition from passive video analysis to active, trustworthy surgical assistance.

2. Results

2.1. Divisions and statistics of the medical center datasets

The datasets used in this study (Cholec80 [20], HeiChole [21], and M2cai16-workflow [22]) are important public resources for surgical video analysis, supporting the development of computer vision models for real-time surgery and robotic-assisted surgery. Each dataset contains annotated surgical videos from multi-medical centers, exhibiting data perspective and institutional heterogeneity that reflect real-world surgical environments. As shown in Fig. 1b, the Cholec80 dataset includes 80 videos, and due to factors such as different camera angles and lighting, the data exhibits certain heterogeneity, such as the proportion of black borders (Fig. 2b). The dataset is divided into three medical centers (medical center 0, medical center 1, and medical center 2). Similarly, the M2cai16-workflow dataset, with 41 video samples, is divided into three medical centers (medical center 5, medical center 6, and medical center 7), focusing on perspective heterogeneity, providing diverse visual information crucial for enhancing model generalization. In contrast, the HeiChole dataset mainly considers institutional heterogeneity, with 24 videos split across two medical centers (medical center 3 and medical center 4). Supplementary Figure 1 visually demonstrates the data heterogeneity across different medical centers by displaying randomly selected consecutive frames from each center. Specifically, differences in equipment, surgical techniques, and procedural practices between medical centers affect the video quality and content, further increasing the

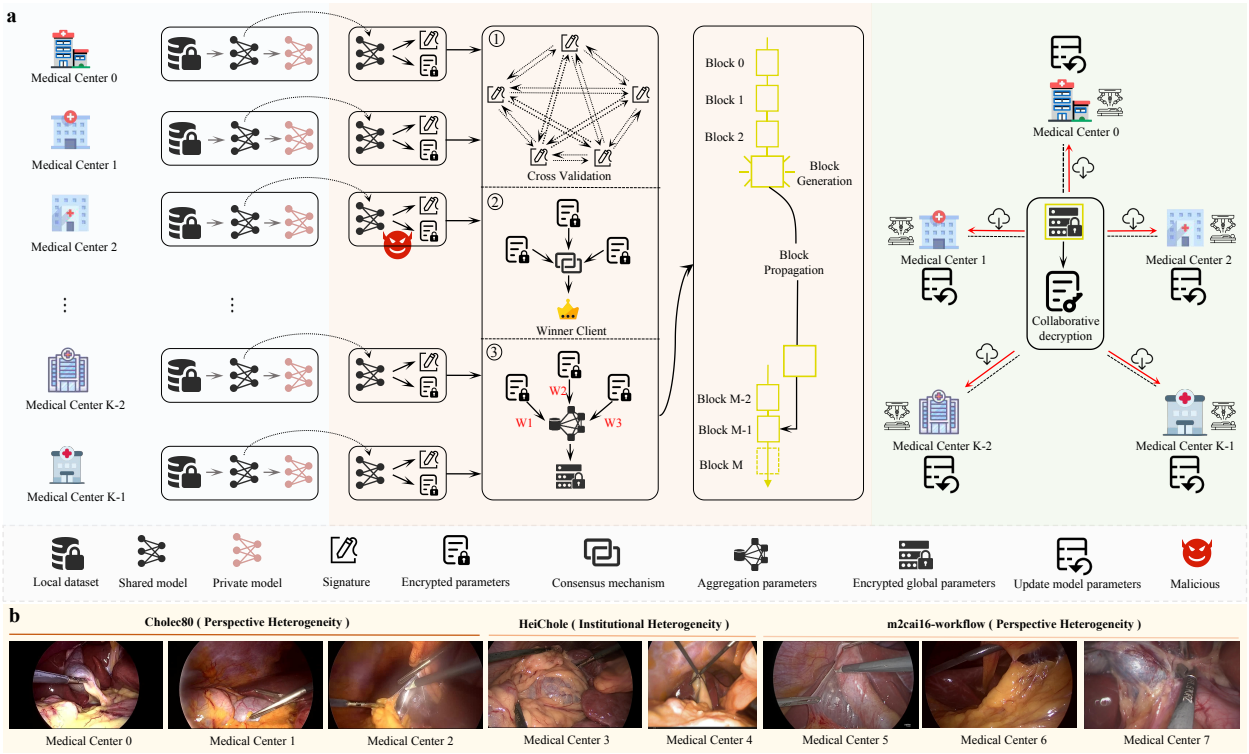


Fig. 2: **The overall architecture of our method**. **a** Overview of the proposed BFL. A collaborative framework for surgical video analysis across multiple medical centers, integrating computer vision and blockchain. The process consists of three main steps: (1) local model training and encryption, (2) cross-validation and block generation, and (3) collaborative parameter decryption among medical centers. **b** Dataset examples highlighting heterogeneity. Frames from Cholec80 (Perspective Heterogeneity), HeiChole (Institutional Heterogeneity), and m2cai16-workflow (Perspective Heterogeneity) datasets illustrate differences in surgical view and environment across medical centers.

complexity and diversity of this dataset. These medical center datasets share a common set of annotated surgical phases, which include preparation, calot triangle dissection, clipping cutting, gallbladder dissection, gallbladder packaging, cleaning coagulation, and gallbladder retraction. Supplementary Figure 2 displays examples of these different surgical phases.

There are significant differences in frame distribution across medical centers, as shown in Fig. 1c. These differences not only reflect the diversity of the data but also realistically simulate the data distribution between different hospitals, further emphasizing the importance of promoting collaboration among multiple medical centers. Additionally, the violin plot depicting the color intensity distribution (RGB channels) in Fig. 1c further illustrates the heterogeneity introduced by different medical centers and datasets. The Cholec80 and M2cai16-workflow datasets exhibit broader distributions in color intensity, particularly in the red and green channels, with variations influenced by lighting conditions, camera equipment, and camera angles. In contrast, the HeiChole dataset shows a more constrained and consistent color intensity distribution.

By incorporating diverse visual perspectives, frame distributions, and other factors, these datasets closely reflect real-world medical environments, providing a valuable benchmark for evaluating the performance and generalizability of computer vision algorithms in surgical phase recognition. The integration of heterogeneous data sources ensures that models trained on these datasets can effectively adapt to real-world clinical surgical settings, helping to promote data sharing and collaboration across different medical centers.

2.2. BFL framework

BFL is a decentralized federated framework designed to provide joint multi-medical center surgical video analysis based on computer vision and blockchain, as shown in Fig. **2**. The process begins with each medical center training its private surgical videos locally. Surgical video segments are encoded using the shared model (ResNet50) [23] and the private model (Temporary Transformer) [24] to understand their content. The shared model extracts important spatial features from the video segments, such as edges, textures, and shapes. Then, the private model processes and integrates the temporal features within the video clips, ensuring that both spatial and temporal features of the video data are effectively explored. The parameters of the shared model are shared across medical centers and contribute to joint training, resulting in a more complete spatial feature encoding model, while the parameters of the private model remain local at each medical center to maintain personalized features. BFL was trained and tested on publicly available surgical datasets to achieve intelligent surgical video analysis. This framework offers a solution to the data island problem in multi-medical-center scenarios, emphasizing data privacy and assisting medical centers in analyzing surgical video.

2.3. Performance evaluation

In this section, we first conduct quantitative comparisons of the proposed method, BFL, with previous studies within the surgical video. We maintain consistent dataset partition settings and evaluation metric calculation methods with the comparative methods. As shown in Fig.**3**, we evaluate the performance of BFL, representative federated learning methods (FedAvg [19], FedProx [25], FedBn [26]), and the Local only method from multiple dimensions. In each subplot, the horizontal axis represents different medical centers (medical center 0 to center 7) and the mean of all medical center results, while the vertical axis denotes the corresponding performance metric values. For a detailed breakdown of the fine-grained precision (PR), recall (RE), F1 score (F1), and Jaccard index (JA) [27] results for each phase of each client, please refer to Supplementary Figures 3-6. By observing the quantitative results of each metric, we can clearly see the advantages of BFL in PR, F1, JA, and RE.

In the comparison of PR results (Fig. **3**a), although the performance of BFL at medical center 3 is comparable to that of FedAvg [19] and FedProx [25], its overall average score for this metric is the highest at 0.832, representing an improvement of 0.020 compared to FedAvg [19]. This indicates that BFL demonstrates superior ability in capturing the temporal and semantic structures of surgical video, providing strong support for surgical video analysis. The F1 score, a metric that balances precision and recall, reflects model performance more comprehensively. As shown in Fig. **3**b, BFL outperforms other methods in all medical

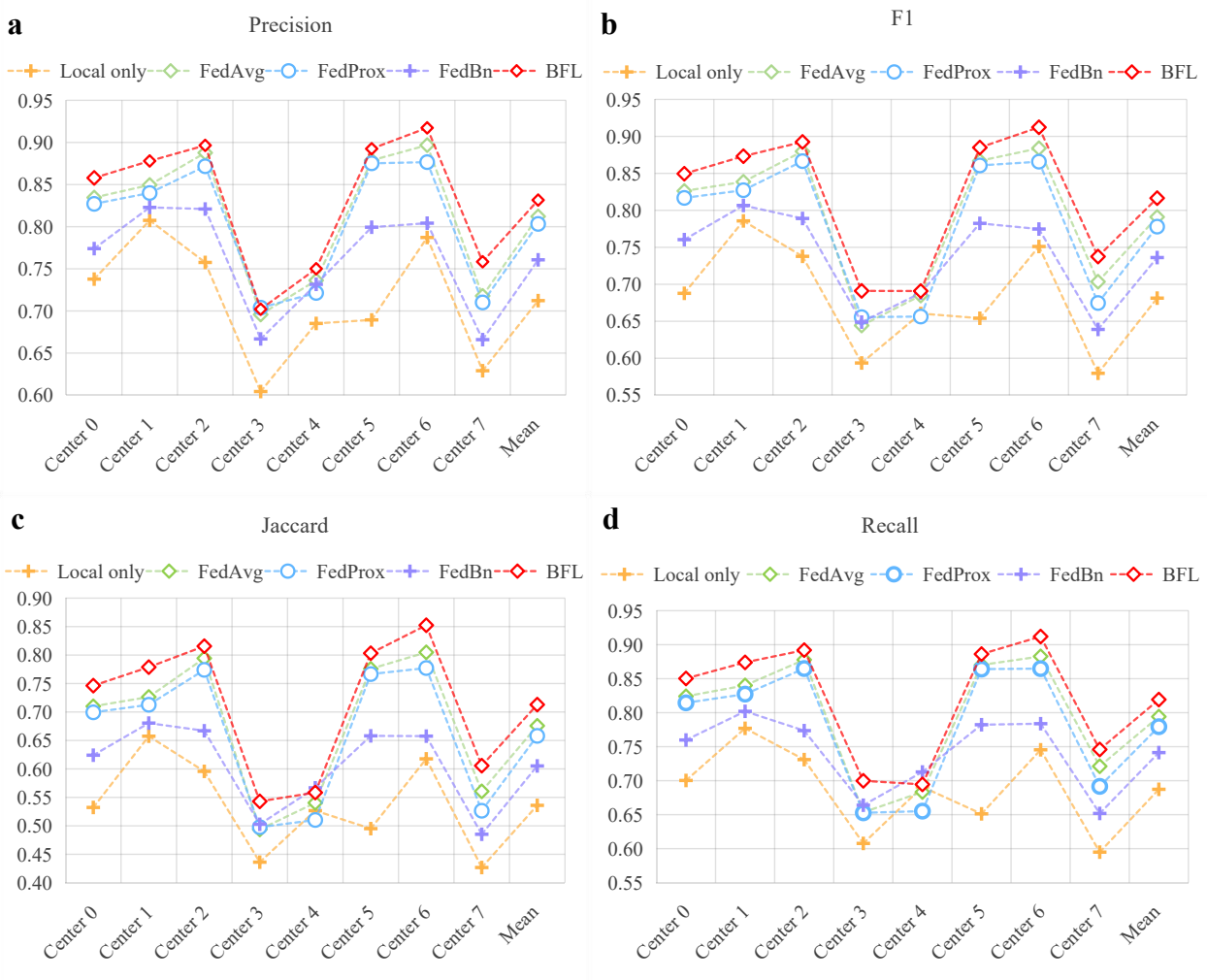


Fig. 3: **Model performance of methods on different medical center**. **a** panel compares the precision of five methods (Local only, FedAvg, FedProx, FedBn, and BFL) across eight medical centers and the mean. Precision evaluates the proportion of correctly predicted positive cases among all predicted positive cases, highlighting the model's ability to avoid false positives. **b** panel shows the F1 score of the methods across centers and the mean. The F1 score balances precision and recall, reflecting the overall effectiveness of the models in handling both false positives and false negatives. **c** panel illustrates the Jaccard index, which measures the overlap between predicted and actual results. It is critical in tasks like segmentation, where accurate spatial correspondence is required. **d** panel presents the recall of the methods, representing the proportion of true positive cases correctly identified. Recall is key in ensuring no true positive cases are missed, particularly in medical diagnosis.

centers. For example, in medical center 1, BFL improves by 0.024 compared to the best alternative method. The average F1 score for BFL is 0.817, significantly higher than the FedAvg [19] score of 0.791. These results demonstrate that BFL has a significant advantage in balancing precision and recall. Another key metric, JA, is used to evaluate the overlap between predicted results and ground truth, which is critical for the accuracy of surgical video analysis. As illustrated in Fig. **3**c, BFL also excels in this metric, achieving an average score of 0.713, representing a notable improvement of 0.037 over FedAvg [19]. This substantial enhancement further supports the applicability of BFL in complex medical scenarios, showing that it not only excels in balancing precision and recall but also in aligning predictions with ground truth labels. Further analysis of recall rate (Fig. **3**d) demonstrates that BFL performs exceptionally well on this metric, highlighting its advantage in coverage capability. At specific medical centers (medical center 0, medical center 1, medical center 5, and medical center 6), BFL consistently outperforms other methods. Although its performance at medical center 4 is slightly inferior to FedBn [26], BFL still maintains an overall advantage in the average recall rate across multiple centers, achieving an average recall of 0.819, representing an improvement of 0.025 compared to the best alternative method (0.794). The results show that BFL demonstrates stronger robustness and generalization ability when faced with complex and variable data distributions in multi-medical center environments, providing solid support for the precise analysis of surgical video.

To further illustrate the performance of our proposed BFL, we present visualizations of confusion matrices and t-SNE plots [28], using medical center 1 and medical center 6 as examples. The confusion matrices (Fig. **4**a and Fig. **4**c) demonstrate that BFL achieves significantly higher accuracy compared to the Local only and FedAvg methods, as evidenced by the notably stronger diagonal values (true positives). This reflects better classification performance. For medical center 1, BFL effectively reduces misclassifications and achieves more accurate predictions across most classes. A similar trend is observed for medical center 6, where BFL consistently minimizes errors compared to the other two methods, highlighting its robustness in handling complex data distributions. The accompanying t-SNE visualizations (Fig. **4**b and Fig. **4**d) further illustrate the superiority of BFL in feature representation. While the Local only and FedAvg methods show scattered and overlapping clusters, BFL produces more compact and well-separated clusters. This indicates that BFL can learn more discriminative and structured features, enhancing its generalization capabilities across different medical centers.

In summary, the combined analysis of confusion matrices and t-SNE visualizations [28] demonstrates that BFL outperforms traditional methods across multiple medical centers. The confusion matrices show that BFL achieves higher true positive rates and fewer misclassifications, while the t-SNE visualizations illustrate its ability to learn compact and well-separated feature clusters. These quantitative and visual analyses effectively highlight the effectiveness of BFL for surgical video analysis. Moreover, the robustness of BFL across heterogeneous institutions and its capacity to extract discriminative representations suggest strong foundational potential for broader applications. The framework is readily extendable to other collaborative medical imaging scenarios, such as radiology, pathology, or endoscopy,

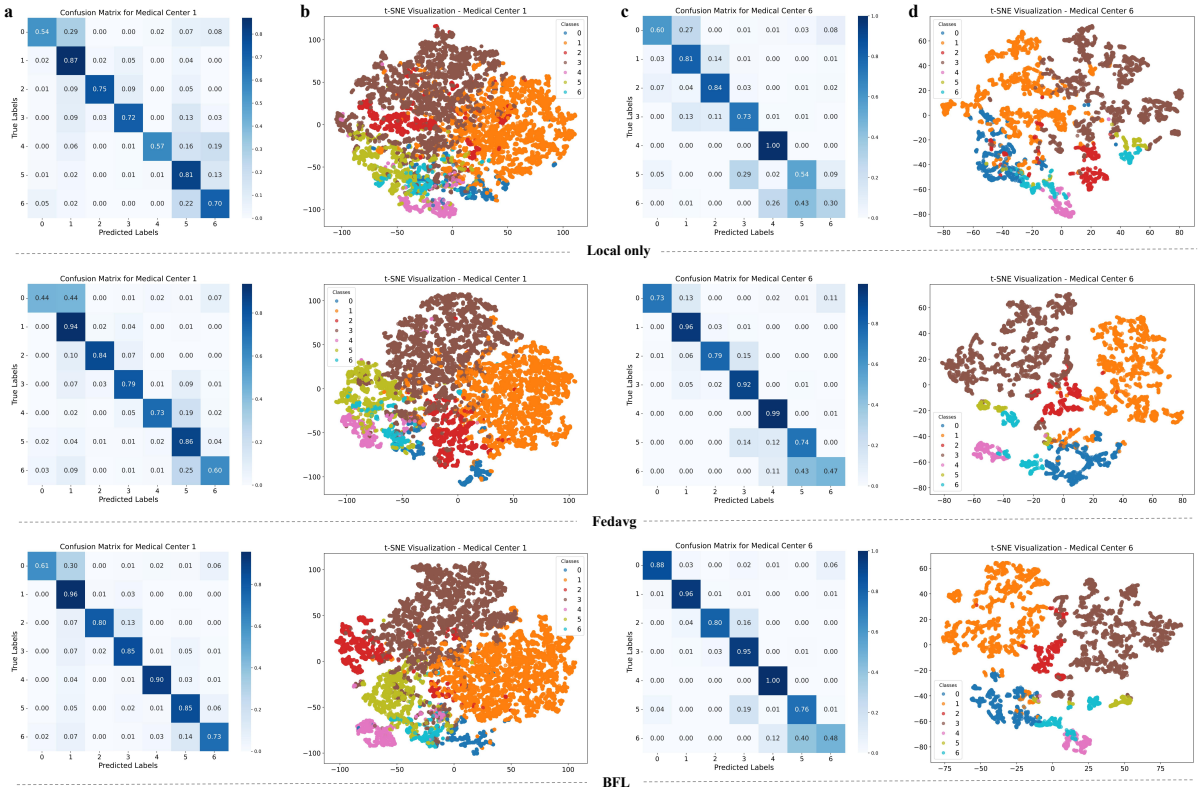


Fig. 4: **Confusion matrices and t-SNE visualizations illustrating model performance.** The figure uses medical center 1 and 6 as examples to demonstrate model classification and feature learning. **a-b** Confusion matrices and t-SNE visualization for medical center 1: Panel (a) shows the confusion matrices for Local only, FedAvg, and BFL, with rows representing true labels and columns representing predicted labels. Panel (b) presents t-SNE visualizations of the feature distributions learned by the three methods, with different colors indicating different classes. **c-d** Confusion matrices and t-SNE visualization for medical center 6: Panel (c) displays the confusion matrices, similar to (a), highlighting the classification results. Panel (d) shows t-SNE plots for the feature embeddings learned by each method, illustrating the clustering of different classes.

where data privacy and institutional diversity are critical concerns. Overall, by establishing a secure mechanism for constructing robust perception capabilities across discrete medical institutions, this study holds promise in supporting future autonomous agents to learn generalizable core skills from diverse real-world data, thereby contributing to the advancement of surgical embodied intelligence.

3. Discussion

In previous studies of federated learning[29, 30, 31], local model parameters from each client were typically uploaded to a central server for unified aggregation. However, this study proposes a novel training strategy within the BFL framework. Unlike traditional methods that involve full-parameter sharing and aggregation, this approach divides the model into two

Table 1: Experimental results validating the effectiveness of the shared and private models design across different medical centers.

Metirc	Method	medical center								
		Center 0	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6	Center 7	Mean
Precision	Base	0.859	0.867	0.900	0.733	0.723	0.886	0.900	0.722	0.824
	Ours	0.858	0.878	0.897	0.702	0.750	0.893	0.917	0.758	0.832
F1	Base	0.853	0.860	0.894	0.674	0.669	0.872	0.892	0.698	0.802
	Ours	0.850	0.873	0.893	0.691	0.691	0.885	0.912	0.738	0.817
Jaccard	Base	0.752	0.759	0.817	0.513	0.521	0.783	0.820	0.554	0.690
	Ours	0.746	0.779	0.816	0.543	0.559	0.803	0.852	0.606	0.713
Recall	Base	0.853	0.861	0.893	0.665	0.669	0.874	0.890	0.714	0.802
	Ours	0.850	0.874	0.892	0.700	0.695	0.886	0.912	0.746	0.819

parts: a shared model and a private model (Fig. **2**a). This design is motivated by the need to balance diversity and personalization in surgical video data, addressing global collaboration and individual requirements in medical scenarios. Analyzing surgical video data necessitates attention to the common features shared across medical centers and each center's unique, personalized features. In this study, the shared model is tasked with extracting spatial features, representing the general visual patterns in surgical videos, such as the appearance of surgical tools and objects within the scene. These features are static by nature and show strong generalization across different centers. In contrast, the private model is designed to extract temporal features that capture dynamic variations unique to each medical center, such as surgical workflows, sequences of tool usage, and other time-dependent characteristics. By modelling spatial and temporal features separately, this approach enhances the generalization capability of the global model while retaining important personalized information.

The experimental results validate the effectiveness of this model separation strategy. As shown in Table 1, the shared and private models (Ours) design demonstrates improvements over the Base approach in metrics such as PR, F1, JA and RE. Through cross-center parameter sharing and aggregation in the shared model, the proposed approach effectively captures and integrates common spatial features of surgical videos, enhancing its robustness and generalization capability. Simultaneously, the localized retention of the private model ensures that each medical center's specific temporal feature requirements are preserved, preventing a decline in the model's adaptability to local data. This design also addresses the issue of information loss caused by parameter aggregation in traditional federated learning, providing a novel solution for personalized feature modelling within federated learning frameworks for medical centers.

To further examine the robustness of the proposed BFL method, we conduct experiments to evaluate its performance against malicious medical centers. Specifically, we analyze how varying numbers of malicious participants impact model accuracy. As shown in Fig **5**, we present the performance of BFL across different metrics (PR, F1, JA, and RE) under

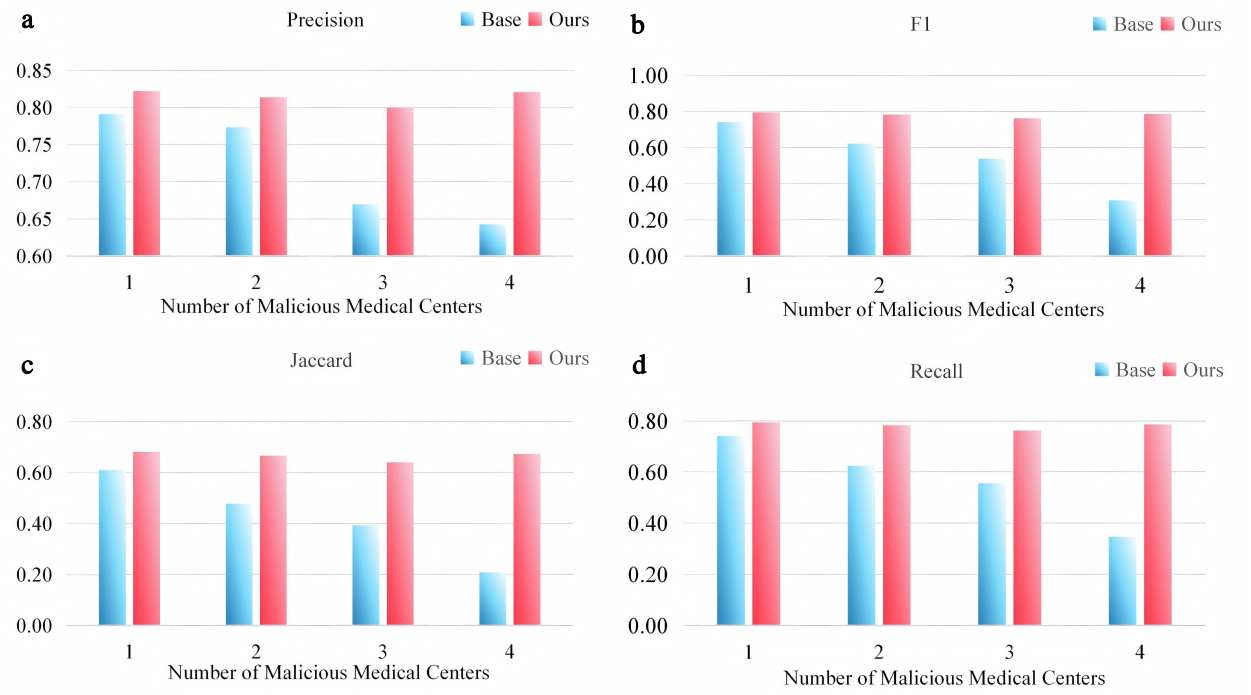


Fig. 5: **Model performance under varying numbers of malicious medical centers. a** panel shows the precision results for the baseline method (Base) and the proposed method (Ours) across different numbers of malicious medical centers, ranging from 1 to 4. Precision evaluates the proportion of correctly predicted positive cases among all predicted positive cases. **b** panel illustrates the F1 score for both methods under different numbers of malicious medical centers. The F1 score reflects the balance between precision and recall. **c** panel presents the Jaccard index, which measures the overlap between the predicted results and ground truth, for different numbers of malicious medical centers. **d** This panel displays the recall for the baseline and proposed methods, indicating the proportion of true positive cases correctly identified as the number of malicious medical centers increases.

varying numbers of malicious medical centers. To ensure consistency in comparisons, we fix the IDs of malicious medical centers for experiments with the same number of malicious participants and measure the average performance metrics only on honest medical centers. The experimental results demonstrate that, regardless of the number of malicious medical centers, BFL maintains high stability across all metrics, further validating its robustness and reliability.

While the benefits of BFL in surgical video analysis have been systematically assessed, this study has several limitations and remaining challenges. Firstly, during the federated training process, the quantity and diversity of surgical data vary across institutions, leading to non-independent and identically distributed (non-IID) data [32, 33]. This situation poses significant challenges to the model aggregation process among different institutions. Although we implemented a balancing strategy that combines data volume and historical performance to adjust the contribution of each institution to the global model, the non-IID problem

persists and may worsen as more institutions engage in the training process. Secondly, the conditions under which public datasets are collected often result in a large portion of the videos coming from the same patient, which limits the variety of features the model can learn. It is anticipated that by expanding the BFL training framework to collaborate with more global medical institutions, while utilizing a more balanced and diverse training dataset, the model's accuracy and generalizability could improve. Lastly, BFL currently does not take advantage of multimodal information. If we could integrate multimodal features, such as combining textual and visual data, the accuracy of the model in surgical video analysis could be further enhanced [34, 35, 36]. Additionally, this integration could facilitate the development of advanced tasks, such as surgical video generation and reconstruction [37]. The present study focuses on surgical phase recognition. However, the progression towards surgical embodied intelligence demands extending these capabilities from passive analysis to active manipulation. Future investigations will therefore aim to utilize the trusted perceptual foundation provided by BFL to support downstream robotic decision-making and control tasks, effectively linking decentralized visual learning with embodied execution[4].

4. Methods

Next, we will present the blockchain-decentralized federated learning model for surgical video analysis (BFL). We first provide the previous definitions of surgical video analysis and the research problem studied in this paper, and then introduce the details of our BFL model.

4.1. Preliminaries and Problem definition

In surgical video analysis, the main task is to recognize the phase P in given video clips $x \in \mathbb{R}^{T\times H\times W\times 3}$. Formally, a set of samples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ is considered, with labels provided as $\mathcal{Y} = \{0, 1, .., C_{P-1}\}$. The goal is to predict a function $f : \mathcal{X} \mapsto \mathcal{Y}$ such that f provides reliable for the unseen data. That is, a given parameterized deep learning model takes the video clips $\boldsymbol{x}$ as input and outputs a set of class-wise presence probabilities for the 7 classes. The objective is to train an intelligent model that minimizes prediction error on the test set, ensuring strong generalization to new, unseen surgical video clips and accurate recognition of the surgical phases [6].

In the context of surgical video analysis, we consider the application of personalized federated learning within a federated setting, involving N different federations, denoted as $\{F_1, F_2, \cdots, F_K\}$, have data, denoted as $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K\}$ with different distributions, which means $P(\mathcal{D}_i) \neq P(\mathcal{D}_j)$. For simplicity, we only study the case where the input and output spaces are the same, i.e. $\mathcal{X}_i = \mathcal{X}_j, \mathcal{Y}_i = \mathcal{Y}_j, \forall i \neq j$. Each dataset, $\mathcal{D}_i = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$, consists of two parts, a train dataset $\mathcal{D}_i^{\text{train}} = \left\{\left(\mathrm{x}_{i,j}^{\text{train}}, y_{i,j}^{\text{train}}\right)\right\}_{j=1}^{\mathrm{n}_i^{\text{train}}}$ and a test dataset $\mathcal{D}_i^{\text{test}} = \left\{\left(\mathbf{x}_{i,j}^{\text{test}}, y_{i,j}^{\text{test}}\right)\right\}_{j=1}^{n_i^{\text{test}}}$. We have $n_i = n_i^{\text{train}} + n_i^{\text{test}}$ and $\mathcal{D}_i = \mathcal{D}_i^{\text{train}} \cup \mathcal{D}_i^{\text{test}}$. We aim to combine information of all federations without data exchange to learn a good model f_i for each federation on its local dataset $\mathcal{D}_i$:

$$\min_{\{f_k\}_{k=1}^{K}} \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i^{\text{test}}} \sum_{j=1}^{n_i^{\text{test}}} \ell \left(f_i \left(\mathbf{x}_{i,j}^{\text{test}} \right), y_{i,j}^{\text{test}} \right). \tag{1}$$

Where ℓ is a loss function. Unlike the previously popular federated learning settings, BFL involves K medical centers without a central server. Each medical center privately keeps its local surgical dataset and does not share it with the outside. The research problem addressed in this paper is to design a decentralized federated learning method for surgical video analysis.

4.2. Details of the model structure of BFL

The BFL framework utilizes the decentralized architecture of blockchain to address single points of failure and trust issues inherent in traditional federated learning central servers. It incorporates both shared and private models across medical centers to enhance collaborative training. By leveraging blockchain's encryption and verification mechanisms, the framework establishes a trusted federated learning environment without compromising sensitive data. The framework consists of three key stages: local model training and encryption, cross-validation and block generation, and collaborative parameter decryption among medical centers.

Each medical center M_k independently maintains its local training dataset D_k and performs training based on it. Before training begins, each center initializes its local shared model θ_0 using a pre-trained shared model architecture, while also initializing a private model δ_k^0 to adapt to its own data characteristics. To ensure data privacy and security, the shared model parameters θ_k^t are encrypted using homomorphic encryption [38], represented as $[\theta_k^t]_{pk_{FHE}}$, allowing them to be used directly in calculations without decryption. Additionally, to protect the integrity and authenticity of the parameters during transmission, each center adds a digital signature $[\theta_k^t]_{signature}$ to the encrypted model parameters, preventing malicious tampering or forgery.

Once the local training is completed and the encrypted parameters $[\theta_k^t]_{pk_{FHE}}$ are uploaded, the system transitions to the cross-validation and block generation phase. During cross-validation, each medical center M_k validates all uploaded shared models $[\theta_n^t]_{pk_{FHE}}$ $(n = 1, 2, \ldots, K)$, starting by confirming the integrity of their digital signatures $[\theta_n^t]_{signature}$ to rule out tampering. The validated models are then tested with local data D_k to assess their performance, summarized in metrics_k, ensuring the parameters' reliability and validity. Only those model parameters that pass cross-validation are deemed trusted and proceed to the next step.

In the consensus phase, the system employs a proof-of-work mechanism to select a winning centerM_n, responsible for aggregating the shared model parameters. This center dynamically allocates aggregation weights based on factors such as historical performance and training costs, resulting in the updated global parameters $[\theta^t]_{pk_{FHE}}$. This dynamic weighting mechanism helps balance both performance fairness and contribution fairness, improving system efficiency and motivating more medical centers to actively engage in model optimization.

Upon successful model aggregation, the winning center generates a candidate block B_t, which includes the aggregated parameters and metadata from participating medical centers (e.g., $[\theta_k^t]_{pk_{FHE}}, [\theta_k^t]_{signature}, \text{metrics}_k$). This block is broadcast to all nodes via the blockchain network. Each node verifies the block's contents upon receipt, and once its legitimacy is confirmed, appends it to the local ledger. Blockchain technology ensures a reliable decentralized architecture, while also providing immutability and transparency for the training results of each round.

Following block generation, the system proceeds to the decryption and update phase for the global shared parameters. medical centers collaborate to decrypt the aggregated shared model $[\theta^t]_{pk_{FHE}}$ using the shared private keys $\{sk_k^{FHE}\}_{k=1}^K$, ensuring that no single node can access the complete decryption information. After decryption, the updated global parameters θ_k^t are downloaded and used to update each local shared model. Through continuous iterations ($t \in [1, e]$), the performance of each model is progressively optimized, while maintaining the privacy of local data.

The entire process integrates homomorphic encryption, digital signatures, cross-validation, dynamic weight aggregation, and blockchain technology to build a secure, fair, and efficient distributed learning framework [39, 40]. This mechanism allows medical centers to collaborate effectively, protect data privacy, jointly improve the global model's performance, and ultimately provide reliable data support and decision-making capabilities for medical applications. We summarize BFL in algorithm 1, and illustrate the details of each process in the following.

Algorithm 1 Workflow of BFL

Input: K medical centers $\{\mathrm{M}_1, \ldots, \mathrm{M}_K\}$ with local training datasets $D = \{D_1, \ldots, D_K\}$, iters e, local learning rate α, local batch size b, initial shared model θ_0, private model δ_k^0.
Output: K medical center model parameters $\eta = \{\eta_1, \ldots, \eta_k\}$.

$\mathrm{M}_1, \ldots, \mathrm{M}_K$ form a blockchain-based network, where M_k has public key pk_k and secret key sk_k.
$\langle sk^{FHE}, pk^{FHE} \rangle$ is generated, then splits sk^{FHE} into K parts, and distributes sk_k^{FHE} to M_k, where $k \in \{1, 2, \ldots, K\}$.
for $t \in [1, e]$ **do**
  if $t \neq 1$ **then**
    $K = N$
  end if
  for each $k \in \{M_1, \ldots, M_K\}$ **do**
    Get shared model $[\theta_{t-1}]_{pk_{FHE}}$ from blockchain;
    $\theta_k^{t-1} \leftarrow$ collab decryption$([\theta_{t-1}]_{pk_{FHE}})$;
    $\eta_k^{t-1} = \theta_k^{t-1} \cup \delta_k^{t-1}$;
    $\theta_k^t, \delta_k^t, \text{metrics}_k \leftarrow$ update local train$(\eta_k^{t-1}, D_k, b, \alpha)$;
    $[\theta_k^t]_{pk_{FHE}}, [\theta_k^t]_{signature} \leftarrow$ encryption and signature$([\theta_k^{t-1}], pk^{FHE}, sk_k)$;
    Broadcast the message $\mathbb{M}_t^k \leftarrow (pk_k, [\theta_k^t]_{pk_{FHE}}, [\theta_k^t]_{signature}, \text{metrics}_k, t)$;
  end for
  Each M_k sends $\mathbb{M}_t^k$ to other medical centers, performs cross-validation on $[\theta_k^t]_{signature}$, and returns the list of validated clients $\{\mathrm{M}_1, \ldots, \mathrm{M}_N\}$ if the verification passes;
  Each M_n abides by the contract mechanism and selects the winning M_n;
  At medical center M_n: Retrieve $\{[\theta_n^t]_{pk_{FHE}}\}_{n=1}^N$ from its memory pool;
  $[\theta^t]_{pk_{FHE}} \leftarrow$ dynamic aggregation$(\{[\theta_n^t]_{pk_{FHE}}\}_{n=1}^N, \text{metrics}_k)$;
  Create $B_t = ([\theta^{t-1}]_{pk_{FHE}}, \mathbb{M}_t^1, \mathbb{M}_t^2, \ldots, \mathbb{M}_t^N)$, and append B_t to the blockchain;
end for
return η_k^e;

4.3. Dataset

In our framework, the shared model used for local training at each medical center is pre-trained on the ImageNet-1K dataset [41]. ImageNet-1K contains over a million images in 1,000 different categories and is widely used for visual recognition tasks. Pre-training on this dataset enables the model to learn rich and diverse visual features, helping it capture complex patterns, textures, and object structures. This enhances the model's ability to generalize, accelerates convergence during fine-tuning, and improves performance on various visual recognition benchmarks.

In this study, we use three challenging laparoscopic cholecystectomy video data sets, divided among 8 medical centers for training and testing. Cholec80 [20] is a large-scale surgical benchmark dataset containing 80 videos with corresponding annotations. These videos were recorded by 13 surgeons during surgeries, with resolutions of either 1920×1080 or 854×480 pixels. HeiChole [21] consists of 24 publicly available laparoscopic cholecystectomy video sequences from Heidelberg University Hospital in Germany. The videos are annotated with labels corresponding to different surgical phases and are primarily used for research on surgical video analysis, focusing on understanding the various phases of the cholecystectomy procedure. M2cai16 [22] includes 41 videos of cholecystectomy surgeries from Strasbourg University Hospital/IRCAD in France and Rechts der Isar Hospital in Munich, Germany. These datasets were originally recorded at 25 frames per second (FPS), but for consistency with previous research [6, 8], we downsampled the frame rate to 1 FPS. Each frame in these surgical videos is annotated with one of seven surgical phases: P1: preparation, P2: calot triangle dissection, P3: clipping and cutting, P4: gallbladder dissection, P5: gallbladder packaging, P6: cleaning and coagulation and P7: gallbladder retraction.

4.4. Metrics

In this paper, we evaluate the performance of BFL using four widely adopted metrics in surgical video analysis: precision (PR), recall (RE), Jaccard index (JA), and F1-score (F1) [27]. These metrics comprehensively assess model performance, addressing key aspects of recognition in surgical video analysis.

PR quantifies the proportion of correctly predicted surgical phases among all predicted phases. High precision minimizes false positives, which is critical in surgical video analysis to avoid mislabeling phases that could lead to erroneous clinical decisions. The formula is:

$$PR = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

where TP and FP denote true positives and false positives, respectively.

RE, also called sensitivity, measures the model's ability to correctly identify all actual instances of a surgical phase. High recall ensures that critical phases, such as gallbladder dissection, are not missed, avoiding incomplete phase recognition. The formula is:

$$RE = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

where FN represents false negatives.

JA evaluates the overlap between predicted and actual phases, balancing precision and recall. It is particularly useful for datasets with imbalanced or overlapping phases, as seen in surgical videos. The formula is:

$$JA = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{4}$$

F1 provides a harmonic mean of precision and recall, offering a single balanced metric. It is especially valuable for handling class imbalance and situations where both false positives and false negatives have significant consequences. The formula is:

$$F1 = 2 \times \frac{PR \times RE}{PR + RE}. \tag{5}$$

These metrics collectively ensure a robust evaluation of the model's performance in recognizing surgical phases.

4.5. Hyperparameters and resources for model training and inference

All experiments are conducted using PyTorch on a single NVIDIA RTX 4090 GPU. We resize all video frames to a uniform dimension of 256×256. The input length of the video clip is set to 6, and the input video frame is enhanced by light data augmentations, such as flipping, rotation, brightness, and saturation perturbations [24, 42, 43]. The specific video IDs used for training and testing at each medical center are detailed in Supplementary Table 1. For local training at each center, we use a shared model based on ResNet50 for visual feature extraction. The shared model is pre-trained on ImageNet-1K [41] to capture rich visual features. The private model at each center is based on a Temporary transformer (T-trans) [24], which learns surgical video-specific temporal features. The hidden dimension of the T-trans model is set to 512. During the training process, we set the batch size to 16 and use an SGD optimizer with an initial learning rate of 1e-2, which decays according to a cosine learning rate schedule. To ensure stable training, we also apply a warmup strategy.

4.6. Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

DATA AVAILABILITY

The dataset used in this study can be queried and downloaded from the following public link:
https://camma.unistra.fr/datasets/
https://www.synapse.org/Synapse:syn25101790/wiki/610856
https://camma.unistra.fr/m2cai2016/

CODE AVAILABILITY

The codes will be available at (https://github.com/halamadrid-lpp/BFL).

FUNDING STATEMENT

The research for this work is supported by the National Natural Science Foundation of China (Grant Nos. 62222207, 61925204, and 62072245).

ACKNOWLEDGEMENT

We are grateful to the CAMMA research group for sharing their expertise in computational analysis and modeling related to medical surgery, which greatly enriched our research experience.

AUTHOR CONTRIBUTIONS

Conceptualization: P. Li., X. Shu., C. M. Feng., and X. Gao. Funding acquisition: X. Shu., and J. Tang. Investigation: P. Li., X. Shu., C. M. Feng., X. Gao. Methodology: P. Li., X. Shu., C. M. Feng. Supervision: X. Shu, C. M. Feng., Y. Feng and J. Tang. Validation: P. Li., X. Shu., Y. Feng and C. M. Feng. Visualization: P. Li., X. Shu., and X. Gao. Writing—original draft: P. Li., X. Shu., and C. M. Feng. Writing—review and editing: All authors

COMPETING INTERESTS

The authors declare no competing interests.

References

[1] Yunlong Li, Zijian Zhao, Renbo Li, and Feng Li. Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. *Artif. Intell. Rev.*, 57:291, 2024.

[2] Pietro Mascagni, Deepak Alapatt, Luca Sestini, Maria S Altieri, Amin Madani, Yusuke Watanabe, Adnan Alseidi, Jay A Redan, Sergio Alfieri, Guido Costamagna, et al. Computer vision in surgery: from potential to clinical value. *NPJ Digit. Med.*, 5:163, 2022.

[3] François Chadebecq, Laurence B Lovat, and Danail Stoyanov. Artificial intelligence and automation in endoscopy and surgery. *Nat. Rev. Gastroenterol. Hepatol.*, 20:171–182, 2023.

[4] Yonghao Long, Anran Lin, Derek Hang Chun Kwok, Lin Zhang, Zhenya Yang, Kejian Shi, Lei Song, Jiawei Fu, Hongbin Lin, Wang Wei, et al. Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics*, 10:eadt3093, 2025.

[5] Recai Yilmaz, Ahmad Alsayegh, Mohamad Bakhaidar, Ali M Fazlollahi, Nour Abou Hamdan, Trisha Tee, Albert Shalmiev, Denis Laroche, Carlo Santaguida, Daniel A Donoho, et al. Combining real-time ai and in-person expert instruction in simulated surgical skills training-randomized crossover trial. *npj Artificial Intelligence*, 1:36, 2025.

[6] Yang Liu, Maxence Boels, Luis C Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Lovit: Long video transformer for surgical phase recognition. *Med. Image Anal.*, 99:103366, 2025.

[7] Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Skit: a fast key information video transformer for online surgical phase recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21074–21084, 2023.

[8] Xinpeng Ding, Xinjian Yan, Zixun Wang, Wei Zhao, Jian Zhuang, Xiaowei Xu, and Xiaomeng Li. Less is more: Surgical phase recognition from timestamp supervision. *IEEE Trans. Med. Imaging*, 42:1897–1910, 2023.

[9] Britty Baby, Daksh Thapar, Mustafa Chasmai, Tamajit Banerjee, Kunal Dargan, Ashish Suri, Subhashis Banerjee, and Chetan Arora. From forks to forceps: A new framework for instance segmentation of surgical instruments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6191–6201, 2023.

[10] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. Fun-sis: A fully unsupervised approach for surgical instrument segmentation. *Med. Image Anal.*, 85:102751, 2023.

[11] Hasan Kassem, Deepak Alapatt, Pietro Mascagni, Alexandros Karargyris, and Nicolas Padoy. Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. *IEEE Trans. Med. Imaging*, 42(7):1920–1931, 2022.

[12] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognit.*, 151:110424, 2024.

[13] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Med. Image Anal.*, 88:102844, 2023.

[14] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.*, 10:12598, 2020.

[15] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention Workshop*, pages 92–104, 2019.

[16] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37:50–60, 2020.

[17] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digit. Med.*, 3:1–7, 2020.

[18] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46:9387–9406, 2024.

[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Machine Learning Research*, pages 1273–1282, 2017.

[20] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging*, 36:86–97, 2016.

[21] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Med. Image Anal.*, 86:102770, 2023.

[22] Andru P. Twinanda, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Single- and multi-task architectures for surgical workflow challenge at m2cai 2016, 2016.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[24] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021.

[25] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

[26] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021.

[27] Shuja Khalid, Mitchell Goldenberg, Teodor Grantcharov, Babak Taati, and Frank Rudzicz. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA network open*, 3:e201664–e201664, 2020.

[28] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nat. Commun.*, 10:5416, 2019.

[29] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nat. Med.*, 27:1735–1743, 2021.

[30] Muhammad Habib ur Rehman, Walter Hugo Lopez Pinaya, Parashkev Nachev, James T Teo, Sebastin Ourselin, and M Jorge Cardoso. Federated learning for medical imaging radiology. *BRIT. J. RADIOL.*, 96:20220890, 2023.

[31] Ahmad Chaddad, Yihang Wu, and Christian Desrosiers. Federated learning for healthcare applications. *IEEE Internet Things J.*, 2023.

[32] Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, and Yan Zhang. Federated learning with non-iid data: A survey. *IEEE Internet Things J.*, 11:19188–19209, 2024.

[33] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594:265–270, 2021.

[34] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *arXiv preprint arXiv:2307.15220*, 2023.

[35] Samuel Schmidgall, Ji Woong Kim, Alan Kuntz, Ahmed Ezzat Ghazi, and Axel Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nat. Mach. Intell.*, 6:1275–1283, 2024.

[36] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 306–316, 2024.

[37] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 230–240, 2024.

[38] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.*, 51:1–35, 2018.

[39] Jing Li, Youliang Tian, Zhou Zhou, Axin Xiang, Shuai Wang, and Jinbo Xiong. Psfl: Ensuring data privacy and model security for federated learning. *IEEE Internet Things J.*, 11:26234–26252, 2024.

[40] KM Sameera, Serena Nicolazzo, Marco Arazzi, Antonino Nocera, Rafidha Rehiman KA, P Vinod, and Mauro Conti. Privacy-preserving in blockchain-based federated learning systems. *Comput. Commun.*, 222:38–67, 2024.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[42] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, et al. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. *Med. Image Anal.*, 86:102803, 2023.

[43] Chinedu Innocent Nwoye, Tong Yu, Saurav Sharma, Aditya Murali, Deepak Alapatt, Armine Vardazaryan, Kun Yuan, Jonas Hajek, Wolfgang Reiter, Amine Yamlahi, et al. Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. *Med. Image Anal.*, 89:102888, 2023.

FIGURE LEGENDS

Fig.1: Comprehensive schematic. **a** Learning strategies: Institutional Incremental Learning (sequential), Cyclic Incremental Learning (iterative refinement), and Decentralized Federated Learning (ours, collaborative with privacy preservation). **b** Dataset summary: Cholec80, HeiChole, and m2cai16-workflow datasets with heterogeneous laparoscopic videos across centers (0–7). **c** Data distribution: (left) Frame counts for training, testing, and total per center; (right) RGB intensity distributions showing visual heterogeneity.

Fig.2: The overall architecture of our method. **a** Decentralized Federated Learning Framework: A collaborative framework for surgical video analysis across multiple medical centers, integrating computer vision and blockchain. The process consists of three main steps: (1) local model training and encryption, (2) cross-validation and block generation, and (3) collaborative parameter decryption among medical centers. **b** Dataset examples highlighting heterogeneity: Frames from Cholec80 (Perspective Heterogeneity), HeiChole (Institutional Heterogeneity), and m2cai16-workflow (Perspective Heterogeneity) datasets illustrate differences in surgical view and environment across medical centers.

Fig.3: Model performance of methods on different medical center. **a** panel compares the precision of five methods (Local only, FedAvg, FedProx, FedBn, and BFL) across eight medical centers and the mean. Precision evaluates the proportion of correctly predicted positive cases among all predicted positive cases, highlighting the model's ability to avoid false positives. **b** panel shows the F1 score of the methods across centers and the mean. The F1 score balances precision and recall, reflecting the overall effectiveness of the models in handling both false positives and false negatives. **c** panel illustrates the Jaccard index, which measures the overlap between predicted and actual results. It is critical in tasks like segmentation, where accurate spatial correspondence is required. **d** panel presents the recall of the methods, representing the proportion of true positive cases correctly identified. Recall is key in ensuring no true positive cases are missed, particularly in medical diagnosis.

Fig.4: Confusion matrices and t-SNE visualizations illustrating model performance. The figure uses medical centers 1 and 6 as examples to demonstrate model classification and feature learning. **a-b** Confusion matrices and t-SNE visualization for medical center 1: Panel (a) shows the confusion matrices for Local only, FedAvg, and BFL, with rows representing true labels and columns representing predicted labels. Panel (b) presents t-SNE visualizations of the feature distributions learned by the three methods, with different colors indicating different classes. **c-d** Confusion matrices and t-SNE visualization for medical center 6: Panel (c) displays the confusion matrices, similar to (a), highlighting the classification results. Panel (d) shows t-SNE plots for the feature embeddings learned by each method, illustrating the clustering of different classes.

Fig.5: Model performance under varying numbers of malicious medical centers. **a** panel shows the precision results for the baseline method (Base) and the proposed method (Ours) across different numbers of malicious medical centers, ranging from 1 to 4. Precision evaluates the proportion of correctly predicted positive cases among all predicted positive cases. **b** panel illustrates the F1 score for both methods under different numbers of malicious medical centers. The F1 score reflects the balance between precision and recall. **c** panel

presents the Jaccard index, which measures the overlap between the predicted results and ground truth, for different numbers of malicious medical centers. **d** This panel displays the recall for the baseline and proposed methods, indicating the proportion of true positive cases correctly identified as the number of malicious medical centers increases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf